1997 Crime Mapping Conference
Exploring the Future of Crime Mapping
Raster-based Analyses

Luc Anselin, West Virginia Regional Research Institute

What is special about spatial data?

1. "Where" matters - the location of points and attributes makes a difference when the data is to be analyzed spatially.
2. Dependence is rule
3. First law of geography - everything is dependent on everything else, but close things are more dependent.

Nature of spatial data:

Spatially referenced data is data that has a geo-referenced attribute associated with the location

There are three different types of spatial objects:

- points (x,y data)
- lines (arcs from point to point)
- polygons (series of connected arcs)

Spatial Processes

Spatial Random Field

- $\{z(s): s \in D\}$
- $s \in R^d$ - a generic data location in a vector of coordinates
- $D \subset R^d$ - the index set which is a subset of the potential locations
- z(s) - a random variable at s with realization z(s)

This list spells out in mathematical notation the requirements for the analysis of some attribute attached to a spatial location. For example, s is a x,y coordinate - z(s) is the land value at that coordinate. z(s) is the realization of the value at location s. It is a random variable because it can take on any value and it is not known ahead of time what the value will be.

An alternate explanation - z(s) is a function that maps some known location s into some value or realization z(s). s is a location that is contained within the space D. D could be any space, but it usually is constrained to some area of interest. s can have several attributes, so it has elements R. For example, the location s can be placed in one, two, three, or four dimensions depending on the study involved.

Classes of Spatial Data (from Cressie)

Point Patterns

D is a point process, s is the realization of that process, e.g. shot location

Geostatistical data

D is a fixed subset of R^d with continual spatial variation, e.g. home sales price

Lattice Data

D is a fixed collection of countably many points in R^d - discrete spaces, e.g. county tax rate districts

[Editor's Note: Another classification of data: (Robinson et al, Elements of Cartography)

- Nominal (qualitative distinction only, no quantitative distinction.)
- Ordinal (differentiation is by class on basis of rank)
- Interval (adds information about distance between ranks)
- Ratio (further refinement of interval, based on non arbitrary zero)]

Spatial Effects

There are two types of spatial effects - spatial heterogeneity and spatial dependence.

Spatial Heterogeneity

- Each location is unique

[Editor's Note: provided the proper scale for the map presentation are used, generalization sometimes will map two uniquely located features to the same geographic space]

- Heterogeneic form is based on the spatial structure

Spatial Dependence

- 2-dimensional and multidirectional
- No future or past for autocorrelation
- Dependent 'sample' contains less information than an independent, identically distributed random sample
- Must test for true contagion vs. apparent contagion

Spatial dependence is a more difficult feature to test for and to use in predictions. Because there is no past and future as with a time series, it is more difficult to use in predictions of influence and value. It is also tricky to test between real and perceived dependencies.

Spatial Association

Null hypothesis: no spatial association

- Values are not dependent on any other values
- Pattern observed is equally likely as any other pattern
- Location of values may be altered without affecting the information content of the data

Hypothesis of Spatial Association

Positive Spatial Association

- Like values cluster in space

- Neighbors are similar-- Value at a location contains info about values at neighboring locations
- Negative Spatial Association
- Dissimilarity of values in neighbors
- Example - Checkerboard pattern - many times this is due to scale problems

Spatial association means, on its basic level, that things that are similar are closer to each other in space. In order to test for spatial association, the 'null hypothesis' must first be rejected. This test determines if a random assignation of values would produce the same result. Because humans are very good at recognizing patterns, even randomly distributed data will appear to have some clustering within it. This is why a mathematical definition is important and a mathematical test is vital. Spatial association can be positive or negative. Usually searches are for positive association, or clustering. However, the negative case can also be important. Negative spatial association gives the perception of a checkerboard - unlike values clustering around like values.

Spatial Data Analysis with GIS - 2 perspectives
- Data driven (exploratory SDA)
- Model driven (spatial econometrics)

There are two types of Spatial Data Analysis according to Dr. Anselin. The first, data driven exploratory Spatial Data Analysis, is used when the data is present and a retrospective analysis or examination of the data is desired. The second, a model driven spatial econometric analysis, is used in more of a predictive fashion. The two areas utilize the same data, and connect at the present, where the observed data ends, and the predicted data begins.

Four Functions of GIS
1. Input
2. Storage
3. Analysis
4. Output

(Functions 1, 2, and 4 will not be discussed during this session)

Number 3 - the Analysis Function
Need to switch back and forth between the GIS module and the spatial data Analysis module in order to refine the selection and analysis

GIS Module
-Selection
-Manipulation
Spatial Data Analysis Module
-Exploration
-Confirmation

What is Exploratory Spatial Data Analysis? The steps are as follows:

- Describe Spatial Distributions
- Identify Outliers - values which do not fit and may skew the analysis
- Discover Patterns of Spatial Association and explain why they are patterns
- Suggest Spatial Regimes
- Identify pockets of local non-stationarity

Note: you must control for spatial dependence

Augment maps with standard exploratory data analysis tools to identify outliers and give a first cut at the data.

Spatial Autocorrelation

{live demos were given of spatial autocorrelation}

Similarity of values corresponds to similarity of location

Mathematical definition of spatial correlation:
$cov[y_i,y_h] \neq 0$ for $i \neq h$

Which i and h interact?
To determine this, we must examine a the values for the covariance. However, if there are N observations, there are N(N-1)/2 interactions, so we must do something intelligent. One such approach is to use a spatial arrangement method. Spatial arrangement imposes a structure on the extent of the spatial interactions. It defines a neighborhood and a spatial weights matrix within that neighborhood. Then, (i,h) pairs are ordered by the distance separation and whether or not they are in the neighborhood for testing.

Another method is the Neighborhood Set of Cliff and Ord (1973), which measures Geo/cartographic contiguity, Spatial Interaction, and Socio-economic distance.

- Geo/cartographic contiguity
  - Boundary
  - Distance
- Spatial Interaction
  - Distance decay, gravity, entropy (scale dependent)
- Socio-economic distance
  - Problem with endogeneity
  - Zero distances

Another neighborhood method is Binary Contiguity Weights. Neighbors are given a value of 1, and non-neighbors are given a value of 0. Neighbors are defined by a common border or vertex.

Binary Contiguity Weights

- NxN positive and symmetric matrix with weights $W_{ij}=1$ for neighbors (0 otherwise)
- W_{ij} where i=j set to 0 by convention
- Contiguity
- Common border, common vertex - islands not always defined

A problem with neighborhood definitions is that neighborhoods are often ambiguous:
Given:

1	2	3
4	5	6
7	8	9

What are the neighbors of 5?

If you are playing chess:
Rook has neighbors 2,4,6,8
Bishop has neighbors 1,3,7,9
Queen has neighbors 1,2,3,4,6,7,8,9

Conclusion: use more than one weight matrix when defining your neighborhoods to ensure your conclusions are supported.

Spatial Lag Operator

- No direct counterpart to time series shift operator ($l_k=y_{t-k}$)
- Usually use weighted average of neighboring values: $\Sigma w_{ij}y_i$ for each i (vector W_y)
- Weights are row standardized $\Sigma w_{ij}=1$

Spatial Autocorrelation statistics:

Moran's I - cross product statistic

$$I = \frac{\left(\frac{N}{S_o}\right)\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

Geary's C - squared difference statistic

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2S_o \sum_i z_i^2}$$

Moran Scatterplot

- Linear Spatial Association
- Linear association between a value at i and a weighted average of neighbors
- $_j w_{ij}y_i$ vs. y_i or w_y vs. y

For more information on how to conduct a Moran Scatterplot, consult a reference on Spatial Statistics.

Types of spatial association

- 4 quats
  - positive association: hi-hi, lo-lo
  - negative association: hi-lo, lo-hi
- Local non-stationarity can be caused by...
  - outliers, high leverage points
  - non-linear association, regimes

Spatial Econometrics or Spatial Regression steps

- Specify structure of spatial dependence
- Test for presence of spatial dependence
- Estimate models with spatial dependence (lag, error, or both)
- Perform Spatial prediction (interpolation, missing values)

Specifying Spatial Dependence

Spatial dependence can be classified in one of two ways: Substantive or Nuisance. The features that each of these categories include are:

- Substantive spatial dependence
  - lag dependence
  - include w_y (neighborhood weight) as explanatory variable
  - $y=\rho w_y+x\beta+\varepsilon$
- Dependence as nuisance (spatial dependence is part of the error or unexplainable features)
  - error dependence
  - non-spherical error variance
  - $E[\varepsilon\varepsilon']=\Omega$, where Ω incorporates the dependence structure

[examples and demonstrations of SpaceStat were given]